Fundamental Limits of Communication With Low Probability of Detection

Ligong Wang, *Member, IEEE*, Gregory W. Wornell, *Fellow, IEEE*, and Lizhong Zheng, *Fellow, IEEE*

***Abstract*— This paper considers the problem of communication over a discrete memoryless channel (DMC) or an additive white Gaussian noise (AWGN) channel subject to the constraint that the probability that an adversary who observes the channel outputs can detect the communication is low. In particular, the relative entropy between the output distributions when a codeword is transmitted and when no input is provided to the channel must be sufficiently small. For a DMC whose output distribution induced by the "off" input symbol is not a mixture of the output distributions induced by other input symbols, it is shown that the maximum amount of information that can be transmitted under this criterion scales like the square root of the blocklength. The same is true for the AWGN channel. Exact expressions for the scaling constant are also derived.**

***Index Terms*— Low probability of detection, covert communication, information-theoretic security, Fisher information.**

I. Introduction

IN MANY secret-communication applications, it is required not only that the adversary should not learn the content of the message being communicated, as in [1], but also that it should not learn whether the legitimate parties are communicating at all or not. Such problems are often referred to as communication with *low probability of detection (LPD)* or *covert communication*. Depending on the application, they can be formulated in various ways.

In [2] the authors consider a wiretap channel model [3], and refer to this LPD requirement as *stealth*. They show that stealth can be achieved without sacrificing communication rate or using an additional secret key. In their scheme, when not sending a message, the transmitter sends some random noise symbols to simulate the distribution of a codeword. There are many scenarios, however, where this cannot be done, because the transmitter must be switched off when not transmitting a message. Indeed, the criterion is often that the adversary should not be able to tell whether the transmitter is on or off, rather than whether it is sending anything meaningful or not. It is the former criterion that is considered in the current paper.

Our work is closely related to the recent works [4]–[6]. In [4] the authors consider the problem of communication over an additive white Gaussian noise (AWGN) channel with the requirement that a wiretapper should not be able to tell with high confidence whether the transmitter is sending a codeword or the all-zero sequence. It is observed that the maximum amount of information that can be transmitted under this requirement scales like the *square root* of the blocklength.[1] In [5] the authors consider a similar problem for the binary symmetric channel and show that the "square-root law" also holds. One major difference between [4] and [5] is that in the former the transmitter and the receiver use a secret key to generate their codebook, whereas in the latter no secret key is used. More recently, Bloch [6] studies the LPD problem from a resolvability perspective and improves upon [4] in terms of secret-key length.

In the current paper, we show that the square-root law holds for a broad class of discrete memoryless channels (DMCs).[2] Furthermore, we provide exact characterizations for the scaling constant of the amount of information with respect to the square root of the blocklength for DMCs as well as AWGN channels, which is not done in [4]–[6].

We do not assume that the eavesdropper observes a noisier channel than the intended receiver; instead, we assume that they both observe the same channel outputs. Our reason for dropping the wiretap structure is that, unlike in secret communication where the assumption that the eavesdropper observes a noisier channel allows one to obtain information-theoretic secrecy without using a secret key, in LPD problems the wiretap assumption does not bring essential new insights. In particular, the square-root law does not rely on the wiretap structure.[3] Hence, by putting the eavesdropper in the same position as the intended receiver, we allow ourselves to focus on the essence of the LPD-communication problem, while at the same time making our results more relevant in practice, the latter because in applications the legitimate parties usually cannot fully determine the statistical behavior of the eaves-

Manuscript received June 10, 2015; revised November 15, 2015; accepted March 14, 2016. Date of publication April 5, 2016; date of current version May 18, 2016. This work was supported in part by NSF within the Division of Computing and Communication Foundations under Grant CCF-1319828 and in part by the Air Force Office of Scientific Research under Grant FA9550-11-1-0183. This paper has been presented in part at the 2015 IEEE International Symposium of Information Theory.

L. Wang is with ETIS, ENSEA/Université de Cergy-Pontoise/CNRS, Cergy-Pontoise 95000, France (e-mail: ligong.wang@ensea.fr).

G. W. Wornell and L. Zheng are with the Department of Electrical Engineering and Computer Science, and the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: gww@mit.edu; lizhong@mit.edu).

Communicated by Y. Liang, Associate Editor for Shannon Theory.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2016.2548471

[1] We adopt the usual terminology to use "blocklength" to refer to the total number of channel uses by a code. However, in the square-root case, the channel codes are not "block codes" in the traditional sense, because they cannot be used repeatedly. Indeed, repeated tramsmission would increase the eavesdropper's probability of detecting the communication.

[2] The achievability part of the square-root law, but not the converse, is independently derived in [6].

[3] In fact, one can verify that the results in [4] hold without the wiretap assumption; see Section V of the current paper for stronger results.

dropper's channel. We also note that extension of most of the results in the paper to wiretap channels is straightforward, part of which can be seen in [6].

Because we do not assume a wiretap structure, contrary to [5], in our setting LPD communication is impossible without a secret key. We assume that such a key is available, and are not concerned with its length within the scope of this paper.

We assume that the receiver does know when the transmitter is sending a message. This is a realistic assumption because the transmitter and the receiver can use part of their secret key to perform synchronization prior to transmission: They choose a (large enough) number of input sequences of a certain length such that each sequence induces an output distribution that is sufficiently different from the output distribution when there is no input to the channel, while on average these sequences induce an output distribution that is sufficiently close to the output distribution when there is no input. Using part of the secret key they randomly pick one of these sequences, which the transmitter sends to the receiver as a synchronization signal before sending a message.

One technical difference between [4] and [5] and the present work is that the earlier works use total variation distance to measure probability of detection whereas we use *relative entropy*, as [2] and [7]. Note that, when the relative entropy is given, the total variation distance can be upper-bounded using Pinsker's inequality [8]. See [2] for further discussions on the relation between relative entropy and detectability. In practice, which of the two quantities is more relevant may depend on the actual application,[4] whereas for theoretical analysis relative entropy is clearly easier to handle.

Summarizing the above discussions, we now briefly describe our setting:

- We consider a DMC whose input alphabet contains an "off" symbol. When the transmitter is switched off, it always sends this symbol.
- The transmitter and the receiver share a secret key that is sufficiently long.
- We assume that the adversary observes the same channel outputs as the intended receiver, i.e., there is no wiretap structure.
- The LPD criterion is that the relative entropy between the output distributions when a codeword is transmitted and when the all-zero sequence is transmitted must be sufficiently small.

The square-root law has been observed in various scenarios in *steganography* [9]–[11]. The setup in steganography that is most related to our work is as follows: a data file called the *cover text* is generated according to some distribution, and a message must be concealed in this file subject to the constraint that the file should look almost unchanged. This is similar to the LPD setting in the sense that, when no message is to be conveyed, the encoder should not do anything, hence, in steganography the output is the original data file, whereas in LPD communications the output is pure noise. But steganography and LPD communications are essentially different: in steganography the data file is generated first and shown to the encoder, whereas in LPD communications noise is added to the codeword after the latter is chosen by the encoder. Hence the two types of problems require different analyses.

The rest of this paper is arranged as follows. In Section II we formulate the problem for DMCs and briefly analyze the case where the "off" input symbol induces an output distribution that can be written as a mixture of the other output distributions; the next two sections focus on the case where it cannot. In Section III we derive formulas for characterizing the maximum amount of information that can be transmitted over any DMC under the LPD constraint. In Section IV we derive a simpler formula that is applicable to some DMCs. In Section V we formulate and solve the problem for AWGN channels. Finally, in Section VI we conclude the paper with some remarks on future directions.

II. Problem Formulation for DMCs

Consider a DMC of finite input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$, and of transition law $W(\cdot|\cdot)$. Throughout this paper, we use the letter P to denote input distributions on $\mathcal{X}$ and the letter Q to denote output distributions on $\mathcal{Y}$. Let $0 \in \mathcal{X}$ be the "off" input symbol; i.e., when the transmitter is not sending a message, it always transmits 0. Denote

$$Q_0(\cdot) \triangleq W(\cdot|0). \tag{1}$$

Without loss of generality, we assume that no two input symbols induce the same output distribution; in particular, $W(\cdot|x) = Q_0(\cdot)$ implies $x = 0$.

A (deterministic) code of blocklength n for message set $\mathcal{M}$ consists of an encoder $\mathcal{M} \to \mathcal{X}^n$, $m \mapsto x^n$ and a decoder $\mathcal{Y}^n \to \mathcal{M}$, $y^n \mapsto \hat{m}$. The transmitter and the receiver choose a *random* code of blocklength n for message set $\mathcal{M}$ using a secret key shared between them. The adversary is assumed to know the distribution according to which the transmitter and the receiver choose the random code, but not their actual choice.[5]

The random code, together with a message M uniformly drawn from $\mathcal{M}$, induces a distribution $Q^n(\cdot)$ on $\mathcal{Y}^n$. We require that, for some constant $\delta > 0$,[6]

$$D\left(Q^n \,\|\, Q_0^{\times n}\right) \le \delta. \tag{2}$$

Here $Q_0^{\times n}$ denotes the n-fold product distribution of Q_0, i.e., the output distribution over n channel uses when the transmitter is off.

At this point, we observe that an input symbol x with $\mathsf{supp}(W(\cdot|x)) \not\subseteq \mathsf{supp}(Q_0)$, where $\mathsf{supp}(\cdot)$ denotes the support of a distribution, should never be used by the transmitter.

[4]The total variation distance would be the right quantity to look at if one assumes equal probabilities for the transmitter sending and not sending a message, because it would correspond to the minimum probability of detection error by the eavesdropper. However, such an assumption is clearly unrealistic in practice.

[5]Note that we assume that the eavesdropper observes the same channel outputs as the intended receiver, so LPD communication is impossible with deterministic codes.

[6]All logarithms in this paper are natural. Accordingly, information is measured in nats.

Indeed, using such an input symbol with nonzero probability would result in $D\left(Q^n \| Q_0^{\times n}\right)$ being infinity. Hence we can drop all such input symbols, as well as all output symbols that do not lie in $\mathsf{supp}(Q_0)$, reducing the channel to one where

$$\mathsf{supp}(Q_0) = \mathcal{Y}. \tag{3}$$

Throughout this paper we assume that (3) is satisfied. Note that, for channels that cannot be reduced to one that satisfies (3), such as the binary erasure channel, nontrivial LPD communication is not possible.

Our goal is to find the maximum possible value for $\log|\mathcal{M}|$ for which a random codebook of length n exists that satisfies condition (2), and whose average probability of error is at most ϵ. (Later we shall require that ϵ be arbitrarily small.) We denote this maximum value by $K_n(\delta, \epsilon)$.

We call an input symbol x *redundant* if $W(\cdot|x)$ can be written as a mixture of the other output distributions, i.e., if

$$W(\cdot|x) \in \mathsf{conv}\left\{W(\cdot|x') \colon x' \in \mathcal{X}, x' \neq x\right\}, \tag{4}$$

where conv denotes the convex hull. As we shall show, $K_n(\delta, \epsilon)$ can increase either linearly with the blocklength n or like $\sqrt{n}$, depending on whether 0 is redundant or not.

A. Case 1: Input Symbol 0 Is Redundant

This is the case where there exists some distribution P on $\mathcal{X}$ such that

$$P(0) = 0 \tag{5a}$$

$$\sum_{x\in\mathcal{X}} P(x)W(\cdot|x) = Q_0(\cdot). \tag{5b}$$

In this case, a positive communication rate can be achieved:

Proposition 1: If input symbol 0 *is redundant, then for any* $\delta \geq 0$,

$$\lim_{\epsilon\downarrow 0}\lim_{n\to\infty}\frac{K_n(\delta,\epsilon)}{n} = \max I(P, W), \tag{6}$$

where the maximum is taken over input distribution P *that satisfies* (5).

Proof: First note that a random codebook generated IID according to P that satisfies (5) yields $D(Q^n\|Q_0^{\times n}) = 0$. By the standard typicality argument [12], when the rate of the code is below $I(P, W)$, the probability of a decoding error can be made arbitrarily small as n goes to infinity. Conversely, for a codebook whose empirical input distribution does not satisfy (5b), $D(Q^n\|Q_0^{\times n})$ grows linearly in n and is hence unbounded as n goes to infinity. Finally, we check that any P that does not satisfy (5a) is suboptimal. Indeed, for any (nontrivial) P that satisfies (5b) but not (5a), let P' be P conditional on $\mathcal{X}\setminus\{0\}$, then P' also satisfies (5b) and $I(P', W) > I(P, W)$. ∎

Example 1: Binary symmetric channel with an additional "off" symbol.

Consider a binary symmetric channel with an additional "off" symbol as shown in Fig. 1. Its optimal input distribution for LPD communication is uniform on $\{-1, 1\}$, and its capacity under the LPD constraint (2) is the same as its capacity without this constraint, and equals $1 - H_b(p)$, where $H_b(\cdot)$ is the binary entropy function.

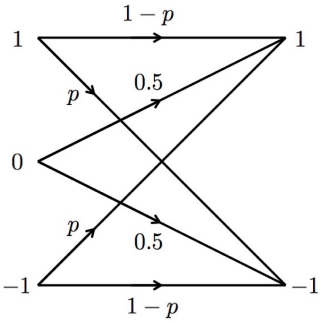

Fig. 1. A binary symmetric channel on the alphabet $\{-1, 1\}$ with cross-over probability p, with an additional "off" input symbol 0 which induces a uniform output distribution.

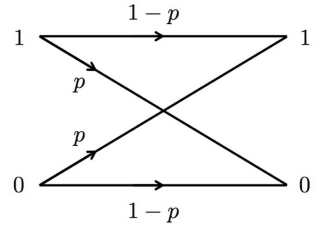

Fig. 2. The binary symmetric channel with cross-over probability p.

B. Case 2: Input Symbol 0 Is Not Redundant

This is the case where no P satisfying (5) can be found. It is the focus of the next two sections. A simple example for this case is the binary symmetric channel in Fig. 2.

We shall show that, in this case, K_n grows like $\sqrt{n}$. Let

$$L \triangleq \lim_{\epsilon\downarrow 0}\varliminf_{n\to\infty}\frac{K_n(\delta,\epsilon)}{\sqrt{n\delta}}, \tag{7}$$

where $\varliminf$ denotes the limit inferior. Note that both $K_n(\delta, \epsilon)$ and δ have unit nat, so L has unit $\sqrt{\text{nat}}$. We shall characterize L in the next two sections. Note that, by definition, L can be infinity, as it is in Case 1.

At this point, we provide some intuition why positive communication rates cannot be achieved in this case. To achieve a positive rate, a necessary condition is that a non-vanishing proportion of input symbols used in the codebook should be different from the "off" symbol 0. This would mean that the average marginal distribution $\bar{P}$ on $\mathcal{X}$ has a positive probability at values other than 0 and, since Q_0 cannot be written as a mixture of output distributions produced by nonzero input symbols, the average output distribution $\bar{Q}$ must be different from Q_0 so $D(\bar{Q}\|Q_0) > 0$. This implies that $D(Q^n\|Q_0^{\times n})$ must grow without bound as n tends to infinity, violating the LPD constraint (2).

III. General Expressions for L for All DMCs

In this section we derive computable expressions for L. Our focus is on Case 2 where 0 is not redundant, though some results also hold (in a trivial way) in Case 1 where 0 is redundant. We first prove the following natural but nontrivial single-letter formula.

Theorem 1: For any DMC,

$$L = \max_{\{P_n\}} \varliminf_{n\to\infty} \sqrt{\frac{n}{\delta}}\, I(P_n, W) \tag{8}$$

where the maximum is taken over sequences of joint distributions on $\mathcal{X} \times \mathcal{Y}$ induced by input distributions P_n and channel W, whose marginals Q_n on $\mathcal{Y}$ satisfy

$$D(Q_n \| Q_0) \le \frac{\delta}{n}. \tag{9}$$

Remark: Although the proof below does not guarantee that the limit inferior in (8) can be replaced by the limit, this is indeed the case, as we show at the end of this section.

Proof of Theorem 1: Proposition 1 shows that, when input symbol 0 is redundant, $L = \infty$. This is consistent with Theorem 1. The rest of the proof focuses on Case 2 as in Section II-B, where 0 is not redundant.

We first prove the converse part. This is done via Fano's inequality and manipulation of the information quantities.

Suppose there exists a sequence of random codes satisfying (2), where, at blocklength n, the size of the codebook is $\exp(K_n)$, and the error probability is ϵ_n which tends to zero as n tends to infinity. By a standard argument using Fano's inequality [13],

$$K_n(1-\epsilon_n) - 1 \le I(X^n; Y^n). \tag{10}$$

Let $\bar{P}_n$ denote the average input distribution on $\mathcal{X}$, averaged over the codebook and over the n channel uses. We upper-bound $I(X^n; Y^n)$ in the usual way:

$$\begin{aligned} I(X^n; Y^n) &= \sum_{i=1}^{n} I(X^n; Y_i | Y^{i-1}) \\ &= \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|X^n, Y^{i-1}) \\ &= \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|X_i) \\ &\le \sum_{i=1}^{n} I(X_i; Y_i) \\ &\le nI(\bar{P}_n, W), \end{aligned} \tag{11}$$

where the last step follows because, when the channel law is fixed, mutual information is concave in the input distribution. Combining (7), (10), and (11) yields

$$L \le \varliminf_{n\to\infty} \sqrt{\frac{n}{\delta}}\, I(\bar{P}_n, W). \tag{12}$$

Next let $\bar{Q}_n$ denote the average output distribution on $\mathcal{Y}$. Clearly, $\bar{Q}_n$ is the output distribution induced by $\bar{P}_n$ through W. Recall that Q^n denotes the n-fold output distribution on $\mathcal{Y}^n$. Further let $Q_{n,i}$ denote the marginal of Q^n on the ith output Y_i. Let Y^n have distribution Q^n, then (see also [14])

$$\begin{aligned} D\left(Q^n \middle\| Q_0^{\times n}\right) &= -H(Y^n) + \mathsf{E}_{Q^n}\left[\log \frac{1}{Q_0^{\times n}(Y^n)}\right] \\ &= -\sum_{i=1}^{n} H(Y_i|Y^{i-1}) + \mathsf{E}_{Q^n}\left[\log \frac{1}{Q_0(Y_i)}\right] \\ &= -\sum_{i=1}^{n} H(Y_i|Y^{i-1}) + \mathsf{E}_{Q_{n,i}}\left[\log \frac{1}{Q_0(Y_i)}\right] \\ &\ge -\sum_{i=1}^{n} H(Y_i) + \mathsf{E}_{Q_{n,i}}\left[\log \frac{1}{Q_0(Y_i)}\right] \\ &= \sum_{i=1}^{n} D(Q_{n,i} \| Q_0) \\ &\ge nD(\bar{Q}_n \| Q_0) \end{aligned} \tag{13}$$

where the last step follows because relative entropy is convex. This combined with (2) implies that

$$D(\bar{Q}_n \| Q_0) \le \frac{\delta}{n}. \tag{14}$$

Combining (12) and (14) proves the converse part of Theorem 1.

We next prove the achievability part. To this end, we randomly generate a codebook that satisfies (2) and then show that, as the length of the codewords tends to infinity, the probability of a decoding error can be made arbitrarily small provided that the codebook has a size smaller than that determined by the right-hand side of (8).

Let $\{P_n\}$ be a sequence of input distributions such that the induced output distributions $\{Q_n\}$ satisfy (9). For every n, we randomly generate a codebook by choosing the codewords IID according to P_n.

It is clear that the output distribution on $\mathcal{Y}^{\times n}$ for this code is $Q^n = Q_n^{\times n}$ and that (2) is satisfied. It remains to show that, provided that the size of the codebook is smaller than $\exp\left(nI(P_n, W) - \sqrt{n}\epsilon_n\right)$ for some ϵ_n tending to zero as n tends to infinity, the probability of a decoding error can be made arbitrarily small. This cannot be shown using the asymptotic equipartition property [12], or the information-spectrum method [15], [16], because we are in a situation where communication rate is zero. However, by slightly varying the methods in [15] and [16], or using the one-shot achievability bounds as in [17] and [18], we can obtain that the sequence $\{K_n\}$ is achievable provided

$$\varlimsup_{n\to\infty} \frac{K_n}{\sqrt{n}} \le P\text{-}\varliminf_{n\to\infty} \frac{1}{\sqrt{n}} \log \frac{W(Y^n|X^n)}{Q_n^{\times n}(Y^n)}, \tag{15}$$

where P-lim inf denotes the *limit inferior in probability*, namely, the largest number such that the probability that the random variable in consideration is greater than this number tends to one as n tends to infinity. Recalling (7), to prove the achievability part of Theorem 1, it now suffices to show that the right-hand side of (15) is lower-bounded by

$$\varliminf_{n\to\infty} \sqrt{n} I(P_n, W).$$

We show a slightly stronger result which is

$$\frac{1}{\sqrt{n}} \log \frac{W(Y^n|X^n)}{Q_n^{\times n}(Y^n)} - \sqrt{n}\, I(P_n, W) \to 0 \quad \text{in probability} \tag{16}$$

as n tends to infinity. To this end, first note

$$\mathsf{E}\left[\frac{1}{\sqrt{n}} \log \frac{W(Y^n|X^n)}{Q_n^{\times n}(Y^n)}\right] = \frac{1}{\sqrt{n}} I(X^n; Y^n) = \sqrt{n}\, I(P_n, W). \tag{17}$$

It then follows by Chebyshev's inequality that, for any constant $a > 0$,

$$\mathsf{Pr}\left[\left|\frac{1}{\sqrt{n}} \log \frac{W(Y^n|X^n)}{Q_n^{\times n}(Y^n)} - \sqrt{n}\, I(P_n, W)\right| \geq a\right] \leq \frac{1}{a^2} \mathsf{var}\left(\frac{1}{\sqrt{n}} \log \frac{W(Y^n|X^n)}{Q_n^{\times n}(Y^n)}\right). \tag{18}$$

Thus, to prove (16), it suffices to show

$$\mathsf{var}\left(\frac{1}{\sqrt{n}} \log \frac{W(Y^n|X^n)}{Q_n^{\times n}(Y^n)}\right) \to 0 \tag{19}$$

as n tends to infinity. To show (19), we first simplify this variance to

$$\mathsf{var}\left(\frac{1}{\sqrt{n}} \log \frac{W(Y^n|X^n)}{Q_n^{\times n}(Y^n)}\right) = \frac{1}{n} \sum_{i=1}^{n} \mathsf{var}\left(\log \frac{W(Y_i|X_i)}{Q_n(Y_i)}\right) = \mathsf{var}\left(\log \frac{W(Y|X)}{Q_n(Y)}\right). \tag{20}$$

The variance on the right-hand side of (20) is upper-bounded by the second moment:

$$\begin{aligned} &\mathsf{var}\left(\log \frac{W(Y|X)}{Q_n(Y)}\right) \\ &\leq \mathsf{E}_{P_n \circ W}\left[\left(\log \frac{W(Y|X)}{Q_n(Y)}\right)^2\right] \\ &= P_n(0)\, \mathsf{E}_{Q_0}\left[\left(\log \frac{Q_0(Y)}{Q_n(Y)}\right)^2\right] \\ &\quad + \sum_{x \neq 0} P_n(x)\, \mathsf{E}_{W(\cdot|x)}\left[\left(\log \frac{W(Y|x)}{Q_n(Y)}\right)^2\right]. \end{aligned} \tag{21}$$

Here we use $P_n \circ W$ to denote the joint distribution on $\mathcal{X} \times \mathcal{Y}$ induced by input distribution P_n through channel W. To prove (19), it suffices to show that both terms on the right-hand side of (21) tend to zero as n tends to infinity. For the first term, note that (9) requires that

$$Q_n \to Q_0 \tag{22}$$

as n tends to infinity, so

$$\lim_{n \to \infty} \log \frac{Q_0(y)}{Q_n(y)} = 0, \quad \forall y \in \mathcal{Y}, \tag{23}$$

which further implies (recall that $|\mathcal{Y}|$ is finite so one can switch the order of limit and expectation)

$$\lim_{n \to \infty} \mathsf{E}_{Q_0}\left[\left(\log \frac{Q_0(Y)}{Q_n(Y)}\right)^2\right] = 0. \tag{24}$$

Thus, since $P_n(0)$ is bounded between 0 and 1, the first term on the right-hand side of (21) tends to zero as n tends to infinity. To analyze the second term on the right-hand side of (21), recall our assumption that Q_0 cannot be written as a mixture of the other output distributions. Thus, to have (22) we need

$$\lim_{n \to \infty} P_n(0) = 1, \tag{25}$$

so

$$\lim_{n \to \infty} P_n(x) = 0, \quad \forall x \neq 0. \tag{26}$$

We next use (22) to obtain (recall again that $|\mathcal{Y}|$ is finite)

$$\lim_{n \to \infty} \mathsf{E}_{W(\cdot|x)}\left[\left(\log \frac{W(Y|x)}{Q_n(Y)}\right)^2\right] = \mathsf{E}_{W(\cdot|x)}\left[\left(\log \frac{W(Y|x)}{Q_0(Y)}\right)^2\right], \tag{27}$$

which is finite for every $x \in \mathcal{X}$, $x \neq 0$, because $Q_0(y) > 0$ for every $y \in \mathcal{Y}$; recall (3). This combined with (26) implies that the second term on the right-hand side of (21) tends to zero as n tends to infinity.

We have now established that the right-hand side of (21) tends to zero as n tends to infinity, which further establishes (19) and, hence, (16). This concludes the achievability part of Theorem 1. ∎

Using Theorem 1 we derive the following computable expression for L.

Theorem 2: For any DMC satisfying (3)*, whose "off" input symbol* 0 *is not redundant, and which has at least one input symbol other than* 0,[7] *L is positive and finite, and is given by*

$$L = \max_{\tilde{P}:\, \tilde{P}(0)=0} \frac{\sum_{x \in \mathcal{X}} \tilde{P}(x) D\left((W(\cdot|x) \| Q_0\right)}{\sqrt{\frac{1}{2} \sum_{y \in \mathcal{Y}} \frac{\left(\tilde{Q}(y) - Q_0(y)\right)^2}{Q_0(y)}}}, \tag{28}$$

where $\tilde{Q}$ is the output distribution induced by $\tilde{P}$ through W.

Before proving Theorem 2 we note that, for some channels, such as the next example, (28) is very easy to compute.

Example 2: Binary symmetric channel.

Consider the binary symmetric channel in Fig. 2. Clearly, the only possible choice for $\tilde{P}$ in (28) is $\tilde{P}(1) = 1$. We thus obtain the value of L as a function of p, which we plot in Fig. 3. Not surprisingly, when p approaches 0.5, L approaches zero, as does the capacity of the channel. It is however interesting to notice that, when p approaches zero, L also approaches zero, even though the capacity of the channel approaches 1 bit per use. This is because, when p is very small, it is very easy to distinguish the two input symbols 0 and 1 at the receiver end. Hence the LPD criterion requires that the transmitter must use 1 very sparsely, limiting the number of information bits it can send. The maximum of L is approximately 0.94 $\sqrt{\text{nat}}$, achieved at $p = 0.083$.

[7] By our assumption, this input symbol induces an output distribution that is different from Q_0, so the channel is not trivial.

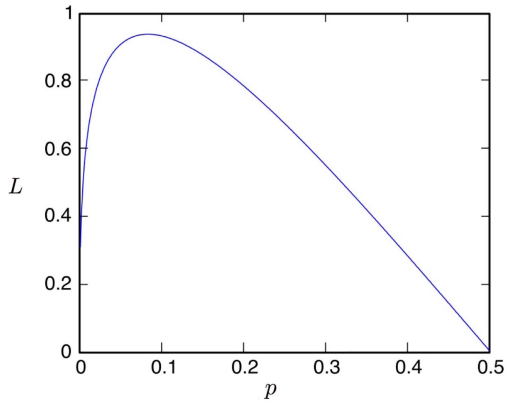

Fig. 3. The value of L for the binary symmetric channel in Fig. 2 as a function of p.

Proof of Theorem 2: For every n, let

$$\hat{P}_n \triangleq \operatorname*{argmax}_{P_n} I(P_n, W) \tag{29}$$

subject to

$$D(Q_n \| Q_0) \leq \frac{\delta}{n}. \tag{30}$$

Using the same argument as for (25), we have

$$\lim_{n\to\infty} \hat{P}_n(0) = 1, \tag{31}$$

hence $\hat{P}_n$ can be written as

$$\hat{P}_n = (1-\mu_n)P_0 + \mu_n \tilde{P}_n \tag{32}$$

where P_0 is the deterministic distribution with $P_0(0) = 1$, $\tilde{P}_n$ is a distribution with $\tilde{P}_n(0) = 0$, and μ_n is positive and tends to zero as n tends to infinity. Fix $\tilde{P}_n$ and consider $\hat{P}_n$ given by (32) as a function of μ_n, then

$$\left.\frac{\mathrm{d}I(\hat{P}_n, W)}{\mathrm{d}\mu_n}\right|_{\mu_n=0} = \sum_{x\in\mathcal{X}} \tilde{P}_n(x) D(W(\cdot|x)\|Q_0), \tag{33}$$

hence

$$I(\hat{P}_n, W) = \mu_n \sum_{x\in\mathcal{X}} \tilde{P}_n(x) D(W(\cdot|x)\|Q_0) + o(\mu_n), \tag{34}$$

where the term $o(\mu_n)$ tends to zero faster than μ_n as n tends to infinity.

The output distribution resulting from feeding $\hat{P}_n$ given by (32) into the channel W is

$$\hat{Q}_n = (1-\mu_n)Q_0 + \mu_n \tilde{Q}_n \tag{35}$$

where $\tilde{Q}_n$ is the output distribution induced by input distribution $\tilde{P}_n$ through W. The relative entropy $D(\hat{Q}_n\|Q_0)$ is approximated by the Fisher Information [19] with respect to parameter μ_n:

$$D(\hat{Q}_n\|Q_0) = \frac{\mu_n^2}{2} \sum_{y\in\mathcal{Y}} \frac{\left(\tilde{Q}_n(y) - Q_0(y)\right)^2}{Q_0(y)} + o(\mu_n^2), \tag{36}$$

where the term $o(\mu_n^2)$ tends to zero faster than μ_n^2 as n tends to infinity. By (30) and (36), μ_n should have the form

$$\mu_n = \sqrt{\frac{\delta}{n}} \cdot \frac{1}{\sqrt{\frac{1}{2}\sum_{y\in\mathcal{Y}} \frac{\left(\tilde{Q}_n(y) - Q_0(y)\right)^2}{Q_0(y)}}} + o\left(n^{-1/2}\right). \tag{37}$$

Plugging (37) into (34) yields

$$I(\hat{P}_n, W) = \sqrt{\frac{\delta}{n}} \cdot \frac{\sum_{x\in\mathcal{X}} \tilde{P}_n(x) D\left((W(\cdot|x)\|\, Q_0\right)}{\sqrt{\frac{1}{2}\sum_{y\in\mathcal{Y}} \frac{\left(\tilde{Q}_n(y) - Q_0(y)\right)^2}{Q_0(y)}}} + o\left(n^{-1/2}\right). \tag{38}$$

When n tends to infinity, $I(\hat{P}_n, W)$ is dominated by the first term on the right-hand side of (38), hence $\tilde{P}_n$ should tend to the (not necessarily unique) distribution that maximizes this term. Recalling Theorem 1, this completes the proof of Theorem 2. ∎

From the proof of Theorem 2 it follows that the limit inferior in (8) can be replaced by the limit, yielding a more convenient expression for L:

Corollary 1: For any DMC,

$$L = \lim_{n\to\infty} \sqrt{\frac{n}{\delta}} \max_{P_n} I(P_n, W) \tag{39}$$

where the maxima are subject to (9).

Proof: We only need to show that the limit in (39) exists. When input symbol 0 is redundant, this limit exists and is infinity. When 0 is not redundant, the proof of Theorem 2 shows that this limit also exists and equals the right-hand side of (28). ∎

IV. A Simpler But Less General Expression for L

In this section we consider channels that satisfy the following condition.

Condition 1: There exists a capacity-achieving input distribution that uses all the input symbols.

Note that Condition 1 implies that no input symbol is redundant; in particular, 0 is not redundant.

We next give a simple upper bound on L under Condition 1. Later we provide an additional condition under which this bound is tight.

Theorem 3: Consider a DMC that satisfies Condition 1. Denote its capacity-achieving output distribution by Q^*, then

$$L \leq \sqrt{2\,\mathsf{var}_{Q_0}\left(\log\frac{Q_0(Y)}{Q^*(Y)}\right)}, \tag{40}$$

where $\mathsf{var}_{Q_0}(\cdot)$ denotes the variance of a function of Y where Y has distribution Q_0.

The proof of Theorem 3 utilizes the following lemma.

Lemma 1: Let Q^* denote the capacity-achieving output distribution for a DMC $W(\cdot|\cdot)$ of capacity C. Let P' be any input distribution, and let Q' denote the output distribution induced by P' through W. Then

$$I(P', W) \leq C - D(Q'\|Q^*), \tag{41}$$

where equality holds if $\mathsf{supp}(P') \subseteq \mathsf{supp}(P^*)$ *for some capacity-achieving input distribution* P^*.

Proof: We have the following identity (see [20]):

$$\begin{aligned} I(P', W) &= \sum_{x\in\mathcal{X}} P'(x) D(W(\cdot|x)\|Q') \\ &= \sum_{x\in\mathcal{X}} P'(x) \mathsf{E}_{W(\cdot|x)}\left[\log \frac{W(Y|x)}{Q'(Y)}\right] \\ &= \sum_{x\in\mathcal{X}} P'(x) \left(\mathsf{E}_{W(\cdot|x)}\left[\log \frac{W(Y|x)}{Q^*(Y)}\right] \right. \\ &\qquad \left. -\mathsf{E}_{W(\cdot|x)}\left[\log \frac{Q'(Y)}{Q^*(Y)}\right]\right) \\ &= \sum_{x\in\mathcal{X}} P'(x) D(W(\cdot|x)\|Q^*) - D(Q'\|Q^*). \end{aligned} \tag{42}$$

By the Kuhn-Tucker conditions for channel capacity [8],

$$D(W(\cdot|x)\|Q^*)) \le C \tag{43}$$

where equality holds if $x \in \mathsf{supp}(P^*)$. We hence have

$$\begin{aligned} C &= \sum_{x\in\mathcal{X}} P^*(x) D(W(\cdot|x)\|Q^*) \\ &\ge \sum_{x\in\mathcal{X}} P'(x) D(W(\cdot|x)\|Q^*), \end{aligned} \tag{44}$$

where equality holds if $\mathsf{supp}(P') \subseteq \mathsf{supp}(P^*)$. Combining (42) and (44) proves the lemma. ∎

Proof of Theorem 3: Since the channel satisfies Condition 1, from Lemma 1 and Corollary 1 we have

$$L = \lim_{n\to\infty} \sqrt{\frac{n}{\delta}} \left(C - \min D(Q_n\|Q^*)\right), \tag{45}$$

where the minimum is over $Q_n \in \mathsf{conv}\{W(\cdot|x)\colon x \in \mathcal{X}\}$ satisfying (9). To determine L, we need to find Q_n that minimizes $D(Q_n\|Q_0)$ for a fixed $D(Q_n\|Q^*)$. To find an upper bound on L, we drop the condition $Q_n \in \mathsf{conv}\{W(\cdot|x)\colon x \in \mathcal{X}\}$ to consider all distributions on $\mathcal{Y}$. Then the minimum is well known to be achieved by a distribution from the exponential family connecting Q_0 and Q^* [21]:

$$Q_n(y) = \frac{Q_0(y)^{1-\lambda_n} Q^*(y)^{\lambda_n}}{\sum_{y'\in\mathcal{Y}} Q_0(y')^{1-\lambda_n} Q^*(y')^{\lambda_n}}, \quad y \in \mathcal{Y} \tag{46}$$

for some $\lambda_n \in [0, 1]$. Indeed, if a distribution Q_n minimizes $D(Q_n\|Q^*)$ for some fixed $D(Q_n\|Q_0)$, then it must minimize

$$(1-\lambda_n) D(Q_n\|Q_0) + \lambda_n D(Q_n\|Q^*)$$

for some $\lambda_n \in [0, 1]$. This sum can be written as

$$\begin{aligned} &(1-\lambda_n) D(Q_n\|Q_0) + \lambda_n D(Q_n\|Q^*) \\ &\quad = D(Q_n\|R_n) - \log \sum_{y'\in\mathcal{Y}} Q_0(y')^{1-\lambda_n} Q^*(y')^{\lambda_n}, \end{aligned} \tag{47}$$

where

$$R_n(y) \triangleq \frac{Q_0(y)^{1-\lambda_n} Q^*(y)^{\lambda_n}}{\sum_{y'\in\mathcal{Y}} Q_0(y')^{1-\lambda_n} Q^*(y')^{\lambda_n}}, \quad y \in \mathcal{Y}. \tag{48}$$

Hence the best choice is $Q_n = R_n$.

It remains to compute $D(Q_n\|Q_0)$ and $D(Q_n\|Q^*)$, where Q_n is of the form (46), for large n. When n is large, Q_n must be close to Q_0 and hence λ_n must be close to zero. In this case, $D(Q_n\|Q_0)$ is approximated by the Fisher Information [19] with respect to parameter λ_n:

$$D(Q_n\|Q_0) = \frac{\lambda_n^2}{2} \mathsf{var}_{Q_0}\left(\log \frac{Q_0(Y)}{Q^*(Y)}\right) + o(\lambda_n^2). \tag{49}$$

This together with the requirement that Q_n must satisfy (9) implies that

$$\lambda_n \le \sqrt{\frac{2\delta}{n \,\mathsf{var}_{Q_0}\left(\log \frac{Q_0(Y)}{Q^*(Y)}\right)}} + o(n^{-1/2}). \tag{50}$$

Next we compute the derivative of $D(Q_n\|Q^*)$, with Q_n given in (46), with respect to λ_n evaluated at $\lambda_n = 0$ to be

$$\left.\frac{\mathrm{d}D(Q_n\|Q^*)}{\mathrm{d}\lambda_n}\right|_{\lambda_n=0} = -\mathsf{var}_{Q_0}\left(\log \frac{Q_0(Y)}{Q^*(Y)}\right). \tag{51}$$

By Condition 1, there exists a capacity-achieving input distribution that uses 0, so

$$\lim_{\lambda_n\downarrow 0} D(Q_n\|Q^*) = D(Q_0\|Q^*) = C. \tag{52}$$

Hence

$$C - D(R_n\|Q^*) = \lambda_n \mathsf{var}_{Q_0}\left(\log \frac{Q_0(Y)}{Q^*(Y)}\right) + o(\lambda_n). \tag{53}$$

Combining (45), (50), and (53) proves (40). ∎

The bound (40) is tight for many channels, e.g., the binary symmetric channel of Example 2. We next provide a sufficient condition for (40) to be tight.

Let $\mathbf{s}$ be the $|\mathcal{Y}|$-dimensional vector given by

$$s(y) = Q_0(y)\left(\log \frac{Q^*(y)}{Q_0(y)} + C\right), \quad y \in \mathcal{Y}. \tag{54}$$

Consider the following system of linear equations with unknowns α_x, $x \in \mathcal{X} \setminus \{0\}$:

$$\sum_{x\in\mathcal{X}\setminus\{0\}} \alpha_x \left(W(\cdot|x) - Q_0\right) = \mathbf{s}. \tag{55}$$

Solving (55) is a simple problem in linear algebra.

Theorem 4: Suppose Condition 1 is satisfied. If (55) *has a nonnegative solution, then* (40) *holds with equality:*

$$L = \sqrt{2\,\mathsf{var}_{Q_0}\left(\log \frac{Q_0(Y)}{Q^*(Y)}\right)}. \tag{56}$$

The intuition behind Theorem 4 is the following: the vector $\mathbf{s}$ represents the tangent of the curve $Q_n(y)$ given by (46) as a function of λ_n at $\lambda_n = 0$. That (55) has a nonnegative solution means that $\mathbf{s}$ lies in the convex cone generated by $\{W(\cdot|x) - Q_0\colon x \in \mathcal{X}\setminus\{0\}\}$. This further implies that, for small enough λ_n, Q_n of the form given by (55) is a valid output distribution, which, as can be seen in the proof of Theorem 3, guarantees (40) to hold with equality. Along a different direction, we provide below a proof utilizing Theorem 2.

Proof of Theorem 4: We use Theorem 2 to prove Theorem 4. Let $\{\alpha_x\colon x \in \mathcal{X}\setminus\{0\}\}$ be a nonnegative solution to (55), and let

$$A \triangleq \sum_{x\in\mathcal{X}\setminus\{0\}} \alpha_x. \tag{57}$$

Then the following constitutes a valid choice for $\tilde{P}$ in (28):

$$\tilde{P}(x) = \frac{\alpha_x}{A}, \quad x \in \mathcal{X} \setminus \{0\}. \tag{58}$$

The corresponding $\tilde{Q}$ is given by

$$\begin{aligned}\tilde{Q} &= \sum_{x \in \mathcal{X}\setminus\{0\}} \frac{\alpha_x}{A} W(\cdot|x) \\ &= Q_0 + \frac{1}{A} \sum_{x \in \mathcal{X}\setminus\{0\}} \alpha_x (W(\cdot|x) - Q_0) \\ &= Q_0 + \frac{\mathbf{s}}{A}. \end{aligned} \tag{59}$$

We evaluate (28) for this choice of $\tilde{P}$ to obtain a lower bound on L. We first compute the denominator, using (59):

$$\begin{aligned}&\sqrt{\frac{1}{2} \sum_{y \in \mathcal{Y}} \frac{\left(\tilde{Q}(y) - Q_0(y)\right)^2}{Q_0(y)}} \\ &\quad = \sqrt{\frac{1}{2A^2} \sum_{y \in \mathcal{Y}} \frac{s(y)^2}{Q_0(y)}} \\ &\quad = \sqrt{\frac{1}{2A^2} \sum_{y \in \mathcal{Y}} Q_0(y) \left(\log \frac{Q^*(y)}{Q_0(y)} + C \right)^2} \\ &\quad = \sqrt{\frac{1}{2A^2} \mathsf{var}_{Q_0} \left(\log \frac{Q^*(Y)}{Q_0(Y)} \right)}. \end{aligned} \tag{60}$$

We next compute the numerator:

$$\begin{aligned}&\sum_{x \in \mathcal{X}\setminus\{0\}} \tilde{P}(x) D\left((W(\cdot|x) \| \, Q_0 \right) \\ &\quad = \sum_{x \in \mathcal{X}\setminus\{0\}} \tilde{P}(x) \sum_{y \in \mathcal{Y}} W(y|x) \log \frac{W(y|x)}{Q^*(y)} \\ &\qquad + \sum_{x \in \mathcal{X}\setminus\{0\}} \tilde{P}(x) \sum_{y \in \mathcal{Y}} W(y|x) \log \frac{Q^*(y)}{Q_0(y)} \\ &\quad = \sum_{x \in \mathcal{X}\setminus\{0\}} \tilde{P}(x) \cdot C + \frac{1}{A} \sum_{\substack{x \in \mathcal{X}\setminus\{0\} \\ y \in \mathcal{Y}}} \alpha_x W(y|x) \log \frac{Q^*(y)}{Q_0(y)} \\ &\quad = C + \frac{1}{A} \sum_{y \in \mathcal{Y}} \log \frac{Q^*(y)}{Q_0(y)} \sum_{x \in \mathcal{X}\setminus\{0\}} \alpha_x W(y|x) \\ &\quad = C + \frac{1}{A} \sum_{y \in \mathcal{Y}} \log \frac{Q^*(y)}{Q_0(y)} \big(A Q_0(y) + s(y) \big) \\ &\quad = C - D(Q_0 \| Q^*) + \frac{1}{A} \sum_{y \in \mathcal{Y}} s(y) \log \frac{Q^*(y)}{Q_0(y)} \\ &\quad = C - C + \frac{1}{A} \sum_{y \in \mathcal{Y}} Q_0(y) \log \frac{Q^*(y)}{Q_0(y)} \left(\log \frac{Q^*(y)}{Q_0(y)} + C \right) \\ &\quad = \frac{1}{A} \mathsf{var}_{Q_0} \left(\log \frac{Q^*(y)}{Q_0(y)} \right), \end{aligned} \tag{61, 62}$$

where (61) follows from (55). Combining Theorem 2, (60), and (62) yields

$$L \geq \sqrt{2 \, \mathsf{var}_{Q_0} \left(\log \frac{Q^*(y)}{Q_0(y)} \right)}. \tag{63}$$

Recalling Theorem 3, both (40) and (63) must hold with equality. ∎

Example 3: A k-ary uniform-error channel.

Consider a channel with $\mathcal{X} = \mathcal{Y} = \{0, 1, \ldots, k-1\}$ and

$$W(y|x) = \begin{cases} 1 - p, & y = x \\ \dfrac{p}{k-1}, & y \neq x \end{cases} \tag{64}$$

where $p \in (0, 1)$. Clearly, its capacity-achieving output distribution Q^* is uniform. It is easy to check that (55) has solution

$$\alpha_x = \frac{p(1-p)\big(\log((k-1)(1-p) - \log p)\big)}{(k-1)(1-p) - p}, \quad x \in \mathcal{X} \setminus \{0\} \tag{65}$$

which is nonnegative. We can hence use Theorem 4 to obtain

$$L = \sqrt{2\upsilon(k, p)} \tag{66}$$

where

$$\begin{aligned}\upsilon(k, p) = (1-p) \left(\log \frac{1}{1-p} \right)^2 + p \left(\log \frac{k-1}{p} \right)^2 \\ - \left((1-p) \log \frac{1}{1-p} + p \log \frac{k-1}{p} \right)^2. \end{aligned} \tag{67}$$

While one might speculate that (56) holds, for example, for all symmetric channels, this is, perhaps surprisingly, not the case. The following example demonstrates this.

Example 4: A ternary symmetric channel.

Consider a ternary symmetric channel where $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$ and

$$W(\cdot|0) = [0.37 \quad 0.01 \quad 0.62] \tag{68a}$$

$$W(\cdot|1) = [0.62 \quad 0.37 \quad 0.01] \tag{68b}$$

$$W(\cdot|2) = [0.01 \quad 0.62 \quad 0.37]. \tag{68c}$$

The right-hand side of (56) yields 0.66 for this channel, but one can check that, in fact, $L = 0.62$. This is because, as Fig. 4 shows, the exponential family connecting Q_0 and Q^* in the neighborhood of Q_0 does not lie in the set of possible output distributions $\mathsf{conv}\{W(\cdot|x) \colon x \in \mathcal{X}\}$, or, roughly equivalently, $\mathbf{s}$ does not lie in the convex cone generated by $\{W(\cdot|x) - Q_0 \colon x \in \mathcal{X} \setminus \{0\}\}$.

V. AWGN Channels

Consider an AWGN channel described by

$$Y = X + Z, \tag{69}$$

where $X \in \mathbb{R}$ is the channel input, $Y \in \mathbb{R}$ is the channel output, and $Z \in \mathbb{R}$ has the zero-mean Gaussian distribution of variance σ^2, denoted $\mathcal{N}(0, \sigma^2)$, and is independent of X. Let the "off" input symbol be 0, so Q_0 is also $\mathcal{N}(0, \sigma^2)$. The encoder and decoder generate a random code as in Section II subject to the LPD constraint (2), and L is again defined as in (7). Note that we do not impose any average- or peak-power

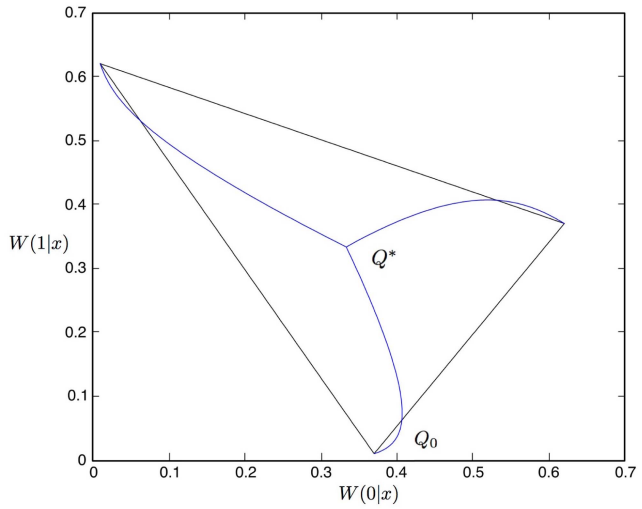

Fig. 4. The ternary symmetric channel in Example 4. The black triangle depicts the set of possible output distributions. The blue curves are the exponential families connecting the conditional output distributions and the capacity-achieving output distribution Q^*. The exponential family connecting Q_0 and Q^* (as the other two exponential families) has a part that lies outside the black triangle, which is why (56) does *not* hold for this channel.

constraint on the input, but imposing such constraints will not affect the value of L due to the stronger LPD constraint (2).[8]

Theorem 5: For an AWGN channel,

$$L = 1 \sqrt{\text{nat}} \tag{70}$$

irrespectively of the noise power σ^2.

The proof of Theorem 5 is divided into the converse part and the achievability part, and is given below.

A. Converse for Theorem 5

Examining the proof of Theorem 1, we see that its converse part is valid for the AWGN channel. Hence

$$L \le \max_{\{P_n\}} \varlimsup_{n\to\infty} \sqrt{\frac{n}{\delta}} I(P_n, W) \tag{71}$$

where the maximum is taken over sequences of joint distributions on $(X, Y) \in \mathbb{R} \times \mathbb{R}$ induced by input distribution P_n via the channel law W resulting from the relation (69), such that the marginal distributions Q_n for Y satisfy

$$D(Q_n \| Q_0) \le \frac{\delta}{n}. \tag{72}$$

Let the second moment of the distribution P_n be denoted ρ_n. It is well known that the zero-mean Gaussian maximizes $I(P_n, W)$ among all distributions of the same second moment (see, e.g., [13]), so

$$I(P_n, W) \le \frac{1}{2} \log \left(1 + \frac{\rho_n}{\sigma^2}\right). \tag{73}$$

[8]The LPD constraint requires that the average input power tend to zero as n tends to infinity, hence rendering any additional average-power constraint inactive. As for peak-power constraints, our choice of input distribution to achieve L is zero-mean Gaussian with vanishing variance. The influence of cutting the tail of such a distribution to meet any peak-power constraint will vanish as n tends to infinity.

Because X and Z are independent, the second moment of the distribution Q_n is $\rho_n + \sigma^2$, yielding

$$\begin{aligned} D(Q_n \| Q_0) &= -h(Q_n) + \mathsf{E}_{Q_n}\left[\log \frac{1}{Q_0(Y)}\right] \\ &= -h(Q_n) + \mathsf{E}_{Q_n}\left[\log\left(\sqrt{2\pi\sigma^2}\, e^{\frac{Y^2}{2\sigma^2}}\right)\right] \\ &= -h(Q_n) + \frac{1}{2}\log\left(2\pi\sigma^2\right) + \mathsf{E}_{Q_n}\left[\frac{Y^2}{2\sigma^2}\right] \\ &= -h(Q_n) + \frac{1}{2}\log\left(2\pi\sigma^2\right) + \frac{\rho_n + \sigma^2}{2\sigma^2}\cdot \\ &\ge -\frac{1}{2}\log\left(2\pi e(\rho_n + \sigma^2)\right) \\ &\quad + \frac{1}{2}\log\left(2\pi\sigma^2\right) + \frac{\rho_n + \sigma^2}{2\sigma^2} \\ &= \frac{\rho_n}{2\sigma^2} - \frac{1}{2}\log\frac{\rho_n + \sigma^2}{\sigma^2}, \end{aligned} \tag{74}$$

where $h(\cdot)$ denotes the differential entropy, and where the inequality follows because the zero-mean Gaussian distribution maximizes differential entropy among all distributions of the same second moment. It follows from (74) that, for $D(Q_n \| Q_0)$ to approach zero as n tends to infinity, ρ_n must tend to zero and

$$D(Q_n \| Q_0) \ge \frac{\rho_n^2}{4\sigma^4} + o(\rho_n^2). \tag{75}$$

Combined with (72), this implies

$$\rho_n \le 2\sigma^2 \sqrt{\frac{\delta}{n}} + o(n^{-1/2}). \tag{76}$$

Plugging this into (73) we obtain

$$\begin{aligned} I(P_n, W) &\le \frac{1}{2}\log\left(1 + \frac{\rho_n}{\sigma^2}\right) \\ &\le \frac{\rho_n}{2\sigma^2} \\ &\le \sqrt{\frac{\delta}{n}} + o(n^{-1/2}). \end{aligned} \tag{77}$$

Combining (71) and (77) yields

$$L \le 1. \tag{78}$$

This concludes the proof of the converse part of Theorem 5.

B. Achievability for Theorem 5

The achievability proof of Theorem 1 relies on the finiteness of the input and output alphabets, therefore it is not applicable to the AWGN channel. Indeed, Theorem 1 may not hold for a general continuous-alphabet channel. However, for the AWGN channel, we only need to prove an achievability result for Gaussian input distributions, which is much simpler than proving it for arbitrary input distributions.

For blocklength n, we randomly generate a codebook such that every codeword is independent of every other codeword, and is IID $\mathcal{N}(0, \rho_n)$ with

$$\rho_n \triangleq 2\sigma^2 \sqrt{\frac{\delta}{n}}. \tag{79}$$

We first check that the LPD condition is met. Indeed, the output sequence is IID $\mathcal{N}(0, \rho_n + \sigma^2)$, so

$$\begin{aligned} D\left(Q^n \,\|\, Q_0^{\times n}\right) &= nD\left(\mathcal{N}\left(0, \rho_n + \sigma^2\right) \|\mathcal{N}\left(0, \sigma^2\right)\right) \\ &= n\left(\frac{\rho_n}{2\sigma^2} - \frac{1}{2}\log\frac{\rho_n + \sigma^2}{\sigma^2}\right) \\ &\le n\left(\frac{\rho_n}{2\sigma^2} - \frac{1}{2}\left(\frac{\rho_n}{\sigma^2} - \frac{\rho_n^2}{2\sigma^4}\right)\right) \\ &= \frac{n\rho_n^2}{4\sigma^4} \\ &= \frac{n}{4\sigma^4}\cdot\left(2\sigma^2\sqrt{\frac{\delta}{n}}\right)^2 \\ &= \delta, \end{aligned} \tag{80}$$

where for the inequality we use the fact

$$\log(1+a) \ge a - \frac{a^2}{2}, \quad a \ge 0. \tag{81}$$

We next look at the maximum number of nats that can be reliably transmitted with this code. Similar to the DMC case, we can show that the sequence $\{K_n\}$ is achievable if (15) holds, except that now Q_n and W are density and conditional density, respectively. The ratio between W and $Q_n^{\times n}$ in (15) can be evaluated as

$$\begin{aligned} \frac{W(y^n|x^n)}{Q_n^{\times n}(y^n)} &= \frac{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - x_i)^2}{2\sigma^2}}}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi(\rho_n+\sigma^2)}} e^{-\frac{y_i^2}{2(\rho_n+\sigma^2)}}} \\ &= \left(\frac{\rho_n + \sigma^2}{\sigma^2}\right)^{\frac{n}{2}} \exp\left(\frac{\sum_{i=1}^n y_i^2}{2(\rho_n+\sigma^2)} - \frac{\sum_{i=1}^n z_i^2}{2\sigma^2}\right). \end{aligned} \tag{82}$$

Hence

$$\begin{aligned} \frac{1}{\sqrt{n}}\log\frac{W(Y^n|X^n)}{Q_n^{\times n}(Y^n)} &= \frac{\sqrt{n}}{2}\log\left(\frac{\rho_n+\sigma^2}{\sigma^2}\right) \\ &\quad + \frac{1}{\sqrt{n}}\left(\frac{\sum_{i=1}^n Y_i^2}{2(\rho_n+\sigma^2)} - \frac{\sum_{i=1}^n Z_i^2}{2\sigma^2}\right). \end{aligned} \tag{83}$$

The mean of (83) satisfies

$$\begin{aligned} &\mathsf{E}\left[\frac{1}{\sqrt{n}}\log\frac{W(Y^n|X^n)}{Q_n^{\times n}(Y^n)}\right] \\ &= \frac{\sqrt{n}}{2}\log\left(\frac{\rho_n+\sigma^2}{\sigma^2}\right) \\ &\quad + \frac{1}{\sqrt{n}}\left(\frac{\sum_{i=1}^n \mathsf{E}\left[Y_i^2\right]}{2(\rho_n+\sigma^2)} - \frac{\sum_{i=1}^n \mathsf{E}\left[Z_i^2\right]}{2\sigma^2}\right) \\ &= \frac{\sqrt{n}}{2}\log\left(\frac{\rho_n+\sigma^2}{\sigma^2}\right) + 0 \\ &\ge \frac{\sqrt{n}}{2}\left(\frac{\rho_n}{\sigma^2} - \frac{\rho_n^2}{2\sigma^4}\right) \\ &= \sqrt{\delta} - \frac{\delta}{\sqrt{n}}, \end{aligned} \tag{84}$$

where we again use (81). By (84) we know that

$$\varliminf_{n\to\infty} \mathsf{E}\left[\frac{1}{\sqrt{n}}\log\frac{W(Y^n|X^n)}{Q_n^{\times n}(Y^n)}\right] \ge \sqrt{\delta}. \tag{85}$$

It remains to show that

$$\lim_{n\to\infty} \mathsf{var}\left(\frac{1}{\sqrt{n}}\log\frac{W(Y^n|X^n)}{Q_n^{\times n}(Y^n)}\right) = 0. \tag{86}$$

Then, by Chebyshev's inequality, we can establish

$$P - \varliminf_{n\to\infty} \frac{1}{\sqrt{n}}\log\frac{W(Y^n|X^n)}{Q_n^{\times n}(Y^n)} \ge \sqrt{\delta} \tag{87}$$

and hence

$$L \ge 1. \tag{88}$$

Using (83), the variance in (86) can be computed as:

$$\begin{aligned} &\mathsf{var}\left(\frac{1}{\sqrt{n}}\log\frac{W(Y^n|X^n)}{Q_n^{\times n}(Y^n)}\right) \\ &= \mathsf{var}\left(\frac{1}{\sqrt{n}}\left(\frac{\sum_{i=1}^n Y_i^2}{2(\rho_n+\sigma^2)} - \frac{\sum_{i=1}^n Z_i^2}{2\sigma^2}\right)\right) \\ &= \frac{1}{n}\sum_{i=1}^n \mathsf{var}\left(\frac{Y_i^2}{2(\rho_n+\sigma^2)} - \frac{Z_i^2}{2\sigma^2}\right) \\ &= \mathsf{var}\left(\frac{Y^2}{2(\rho_n+\sigma^2)} - \frac{Z^2}{2\sigma^2}\right) \\ &= \mathsf{E}\left[\left(\frac{Y^2}{2(\rho_n+\sigma^2)} - \frac{Z^2}{2\sigma^2}\right)^2\right] \\ &= \frac{1}{4(\rho_n+\sigma^2)^2}\mathsf{E}\left[\left(X^2 + 2XZ - \frac{\rho_n}{\sigma^2}Z^2\right)^2\right] \\ &\le \frac{1}{4\sigma^4}\mathsf{E}\left[\left(X^2 + 2XZ - \frac{\rho_n}{\sigma^2}Z^2\right)^2\right]. \end{aligned} \tag{89}$$

After expanding the square inside the expectation in (89), one can verify that the expectation of every summand tends to zero as n tends to infinity, establishing (86), and hence (87) and (88), proving the achievability part of Theorem 5.

VI. Concluding Remarks

A DMC in practice often represents discretization of a continuous-alphabet channel. For example, Figs. 1 and 2 can result from two different discretizations of the same AWGN channel. In this sense, our results suggest that the optimal discretization may depend heavily on whether there is an LPD requirement or not.

In practice, LPD communication systems of positive data rates often can be implemented even when the channel model does not seem to allow positive rates. Indeed, in such applications, the concern is often not that the transmitted signal should be sufficiently weak, but rather that it should have a wide spectrum and resemble white noise [22]. We believe that one of the reasons why such systems may work is that realistic channels often have memory. For example, on a channel whose noise level varies with a coherence time that is longer than the length of a codeword, the transmitter and the receiver can use the adversary's ignorance of the actual noise level to communicate without being detected. One way

to formulate this scenario is to assume that the channel has an unknown parameter that is fixed. This is discussed for the binary symmetric channel in [23]. Further addressing this scenario is part of ongoing research.

ACKNOWLEDGEMENTS

The authors thank Boulat Bash and Matthieu Bloch for helpful comments.

REFERENCES

[1] C. E. Shannon, "Communication theory of secrecy systems," *Bell Syst. Tech. J.*, vol. 28, no. 4, pp. 656–715, 1949.
[2] J. Hou and G. Kramer, "Effective secrecy: Reliability, confusion and stealth," in *Proc. IEEE Int. Symp. Inf. Theory*, Honolulu, HI, USA, Jun./Jul. 2014, pp. 601–605.
[3] A. D. Wyner, "The wire-tap channel," *Bell Syst. Tech. J.*, vol. 54, no. 8, pp. 1355–1387, 1975.
[4] B. A. Bash, D. Goekel, and D. Towsley, "Limits of reliable communication with low probability of detection on AWGN channels," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 1921–1930, Sep. 2013.
[5] P. H. Che, M. Bakshi, and S. Jaggi, "Reliable deniable communication: Hiding messages in noise," in *Proc. IEEE Int. Symp. Inf. Theory*, Istanbul, Turkey, Jul. 2013, pp. 2945–2949.
[6] M. R. Bloch, "Covert communication over noisy channels: A resolvability perspective," *IEEE Trans. Inf. Theory*, to be published.
[7] C. Cachin, "An information-theoretic model for steganography," *Inf. Comput.*, vol. 192, no. 1, pp. 41–56, Jul. 2004.
[8] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. San Diego, CA, USA: Academic, 1981.
[9] A. D. Ker, "A capacity result for batch steganography," *IEEE Signal Process. Lett.*, vol. 14, no. 8, pp. 525–528, Aug. 2007.
[10] J. Fridrich, *Steganography in Digital Media: Principles, Algorithms, and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[11] T. Filler and J. Fridrich, "Fisher information determines capacity of ϵ-secure steganography," in *Information Hiding* (Lecture Notes in Computer Science). Berlin, Germany: Springer-Verlag, 2009, pp. 31–47.
[12] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.
[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991.
[14] J. Hou, "Coding for relay networks and effective secrecy for wire-tap channels," Ph.D. dissertation, Fakultät für Elektrotechnik Informationstechnik, Tech. Univ. München, München, Germany, 2014.
[15] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, Jul. 1994.
[16] T. S. Han, *Information-Spectrum Methods in Information Theory*. Berlin, Germany: Springer-Verlag, 2003.
[17] L. Wang, R. Colbeck, and R. Renner, "Simple channel coding bounds," in *Proc. IEEE Int. Symp. Inf. Theory*, Seoul, South Korea, Jun./Jul. 2009, pp. 1804–1808.
[18] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.
[19] S. Kullback, *Information Theory and Statistics*. New York, NY, USA: Wiley, 1959.
[20] F. Topsøe, "An information theoretical identity and a problem involving capacity," *Stud. Sci. Math. Hung.*, vol. 2, pp. 291–292, 1967.
[21] I. Csiszár and F. Matúš, "Information projections revisited," *IEEE Trans. Inf. Theory*, vol. 49, no. 6, pp. 1474–1490, Jun. 2003.
[22] M. Simon, J. Omura, R. A. Scholtz, and B. K. Levitt, *Spread Spectrum Communications Handbook*. New York, NY, USA: McGraw-Hill, 1994.
[23] P. H. Che, M. Bakshi, C. Chan, and S. Jaggi, "Reliable deniable communication with channel uncertainty," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Hobart, TAS, Australia, Nov. 2014, pp. 30–34.

Ligong Wang (S'08–M'12) received the B.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 2004, and the M.Sc. and Dr.Sc. degrees in electrical engineering from ETH Zurich, Switzerland, in 2006 and 2011, respectively. In the years 2011-2014 he was a Postdoctoral Associate at the Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology, Cambridge, MA, USA. He is now a researcher (chargé de recherche) with CNRS, France, and is affiliated with ETIS laboratory in Cergy-Pontoise. His research interests include classical and quantum information theory, and digital, in particular optical communications.

Gregory W. Wornell (S'83–M'91–SM'00–F'04) received the B.A.Sc. degree in electrical engineering from the University of British Columbia, Vancouver, BC, Canada, and the S.M. and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 1985, 1987, and 1991, respectively.

Since 1991, he has been on the faculty at MIT, where he is the Sumitomo Professor of Engineering in the department of Electrical Engineering and Computer Science (EECS). He leads the Signals, Information, and Algorithms Laboratory in the Research Laboratory of Electronics, and co-chairs the EECS department graduate program. He has held visiting appointments at the former AT&T Bell Laboratories, Murray Hill, NJ, USA, the University of California, Berkeley, CA, USA, and Hewlett-Packard Laboratories, Palo Alto, CA, USA. His research interests and publications span the areas of signal processing, digital communication, and information theory, and include algorithms and architectures for wireless networks, sensing and imaging systems, digitally enhanced analog circuits and systems, multimedia applications, and aspects of computational biology and neuroscience.

Dr. Wornell has been involved in the Information Theory and Signal Processing Societies of the IEEE in a variety of capacities, and maintains a number of close industrial relationships and activities. He has won a number of awards for both his research and teaching.

Lizhong Zheng (S'00–M'02–F'15) received the B.S and M.S. degrees in 1994 and 1997, respectively, from the Department of Electronic Engineering, Tsinghua University, Beijing, China, and the Ph.D. degree in 2002, from the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA. Since 2002, he has been working in the Department of Electrical Engineering and Computer Sciences, where he is currently a professor of Electrical Engineering and Computer Sciences. His research interests include information theory, wireless communications, and statistical inference. He received the IEEE Information Theory Society Paper Award in 2003, and NSF CAREER award in 2004, and the AFOSR Young Investigator Award in 2007. He served in the years 2008–2011 as Associate Editor for Communications for the IEEE TRANSACTIONS ON INFORMATION THEORY.